ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 2 - 2026

Optimizing Business English Writing Scoring and Personalized Feedback Model Using Machine Learning

Teng Mu[1,*]

[1] Foreign Language School, College of Arts and Science of Hubei Normal University, Huangshi 435000, Hubei, China

SUMMARY: *This study develops a machine learning approach to score business English writing and generate feedback that is tied to the actual text. Existing automated scoring systems are usually reliable for surface features such as grammar, vocabulary, and sentence form, but they are less consistent when the task involves tone, coherence, or how well a response fits a business context. In many cases, the feedback they produce is also too broad to help with revision. To address this problem, we organize the model into three parts. The first part learns a representation of writing quality so that responses with similar wording but different communicative effectiveness can be separated more clearly. The second part segments the text and examines local structure, including sentence-level variation and the progression of ideas. The third part ranks candidate feedback and removes comments that are repetitive or weakly connected to the relevant passage. The three parts are used jointly during scoring and feedback generation, rather than as separate stages with little interaction. This reduces overlap between modules and helps keep the model stable across different writing samples. In the experiments, the model performed better than the baselines on both scoring and feedback. Correlation with human ratings increased from 82% to 89%, and feedback precision rose from 71% to 86%. User satisfaction also improved from 3.6 to 4.4 out of 5, suggesting that the revised feedback was more specific and easier to use in practice.*

KEYWORDS: *business English writing; automated writing assessment; personalized feedback; discourse analysis; writing quality evaluation*

1 Introduction

In international business settings, English is used extensively in emails, reports, and day-to-day cross-border communication[1]. The quality of this writing affects not only how clearly a message is delivered, but also how it is interpreted and acted upon[2]. Yet evaluating business English writing remains difficult[3]. Human scoring can vary across raters, even when the texts are similar, and the time required for manual assessment makes large-scale use difficult[4]. At the same time, many learners now expect feedback that points to specific weaknesses in their writing rather than broad comments that are hard to apply in revision[5]. This gap between current assessment practice and learner expectations makes it necessary to improve both scoring reliability and the usefulness of feedback[6]. The issue is not only technical; it also matters for the quality of communication in professional contexts[7].

Early writing assessments relied on rule-based systems[8]. These systems scored texts based on predefined rules of grammar, vocabulary, and sentence structure. But when writing

*lenadsada46456mou@163.com
https://doi.org/10.65102/is2026535
Published: 30 April, 2026

changes in terms of style, purpose, or context, they become less practical[9]. When the writing style or context changes, their performance often drops significantly. Building and maintaining these rules requires continuous expert input, which limits scalability[10]. Even so, this work clearly evaluated which linguistic features are important, opening a starting point for the development of subsequent models[11]. Statistical methods were introduced, because rule-based scoring has poor generalization effect in different writings[12]. These methods, unlike predefined rules, will use features such as word distribution, syntactic patterns and semantic similarity to perform data learning[13]. This makes them more flexible in different tasks and more stable compared with early systems[14]. But this improvement is limited, their performance still depends on manual feature design and preprocessing, so important parts of writing are often not modeled well[15]. Text organization, tone shift and other situations, as well as whether the response is suitable for business environment and other conditions, especially under this circumstance[16]. Neural models, especially transformer-based ones, have changed the game. They model text in context, not as isolated features[17]. They can more efficiently capture inter-sentence relationships, laying a firmer foundation for writing assessment[18]. When there is sufficient training data, they reduce the need for handcrafted features, and can also provide targeted feedback[19]. In commercial writing tasks, fine-tuning for the task can often improve results again. However, these models are not very easy to use in practice. The calculation cost is relatively high; the actual system needs to make a trade-off between performance and training deployment restrictions[20].

To address these problems, we develop a model for scoring and feedback in business English writing. It builds on neural methods, but the design keeps computation relatively manageable and focuses on feedback that can be used directly in revision. The model combines structured representation, segment-level analysis, and probabilistic feedback filtering. In our setting, the same system can be used for emails, reports, and proposal-based writing with only limited adjustment. It also generates feedback from local parts of the text rather than from the document-level score alone. The main contributions of this work are as follows:

- We build a scoring and feedback model that combines neural methods with a relatively compact architecture, with the aim of improving performance without making the system too expensive to use.
- The model is not limited to one business writing task. In our experiments, it transfers to emails, reports, and proposal-based writing with only minor changes.
- Experimental results show improvements over the baselines in both scoring consistency and feedback precision.

2 Method

2.1 Overview

This section describes the structure of the model used for scoring and feedback. It includes three parts: the basic task formulation (2.2), the scoring and segmentation model (2.3), and the training strategy (2.4). Each part addresses a different part of the pipeline. Taken together, they are used to improve score stability and make the feedback more usable during revision. The Preparatory Work part builds up the basic problem environment and puts forward the variables that are used afterward in the model. As a special case, the quality of writing is expressed as a latent variable q, and at the same time, the relevance of feedback is marked by r. These variables cannot be directly seen, but are obtained through inference in the process of

both training and inference. This section moreover gives the definition of the notation which is used in the whole paper, hence the later display of the model can keep consistency. The Manifold Event Optimizer (2.3) is the core forming part of our framework. It is gotten up to split writing into smaller units and assess them in a structured way. This model inside contains three modules. The Counterfactual Manifold Shaper builds a hidden space $\mathcal{M}$ which captures both structural and meaning features of the text. This lets the system carry out comparison between one certain writing work and other possible different changing forms. The Agent-driven Event Segmenter which is made by people cuts the text into a ordered series of events $\{e_1, e_2, \dots, e_n\}$, for example sentence level or discourse level units, hence local problems can be respectively checked by us. The Probability-type Feedback Filtration device hence chooses feedback in accordance with the relevance that has been estimated, which is expressed as $P(r \mid e, q)$. In the actual doing, this step can assist get rid of the suggestions that are redundant or have low influence. For the coordination of these components, Section 2.4 brings in a constrained optimization strategy. Instead of only doing optimization for score correctness, the model moreover brings feedback quality and calculation expense into consideration. This is given the form of

$$\max_x f(x) \quad \text{under the condition that} \quad g(x) \le \epsilon,$$

wherein $f(x)$ expresses scoring and feedback effect, and $g(x)$ represents the usage of resources. The restrictive parameter ϵ holds the control of the balance, which lets the system keep enough efficiency for actual application.

2.2 Preliminaries

We treat business English writing assessment as a joint task of scoring and feedback generation. Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ denote the dataset, where x_i is a writing sample and y_i is its score. The scoring model learns a function $f: \mathcal{X} \to \mathcal{Y}$ that predicts the score from the input text. Each sample is represented by a feature vector $\mathbf{x}_i = [x_{i1}, x_{i2}, \dots, x_{im}]$, which includes syntactic, semantic, and stylistic information. These features are modeled jointly because writing quality is not determined by any one of them alone. We use a latent variable z_i to represent the underlying quality of the text and write the observed score as

$$y_i = g(z_i) + \epsilon_i,$$

where $g(\cdot)$ maps latent quality to the observed score, and ϵ_i captures noise from rating variation. The latent variable is inferred from the input features through

$$z_i = h(\mathbf{x}_i).$$

Feedback is generated from the predicted score $\hat{y}_i = f(x_i)$, the reference score y_i, and the input features:

$$\mathbf{f}_i = \phi(y_i, \hat{y}_i, \mathbf{x}_i),$$

where $\phi(\cdot)$ returns feedback linked to specific parts or properties of the text.

The training objective includes both scoring error and feedback quality:

$$\min_{f,\phi} \sum_{i=1}^{N} \mathcal{L}\left(y_i, f(x_i)\right) + \lambda \sum_{i=1}^{N} \mathcal{R}\left(\mathbf{f}_i\right),$$

where $\mathcal{L}(\cdot)$ is the scoring loss, $\mathcal{R}(\cdot)$ is the feedback term, and λ controls their relative weight.

2.3 Manifold Event Optimizer

Figure 1 summarizes the main structure of the model used for scoring and feedback. We do not treat the response as a single block and assign a score in one step. That approach can miss local problems, especially when the text is grammatically acceptable overall but uneven in structure, coherence, or task fit. Instead, the model processes the text in three linked stages. It first maps the response into a structured latent representation, then analyzes smaller units of the text in sequence, and finally generates feedback from the resulting local and global signals. These three stages play different roles. The latent-space component is used to represent writing quality in a more compact form while preserving the main variation in the input features. The segment-level component then examines how different parts of the text contribute to the score, rather than assuming that all parts are equally informative. On top of this, the feedback component ranks candidate comments and filters out suggestions that are repetitive or only weakly related to the segment being evaluated. Taken together, this design links score prediction with feedback generation more directly than a document-level scoring pipeline alone.

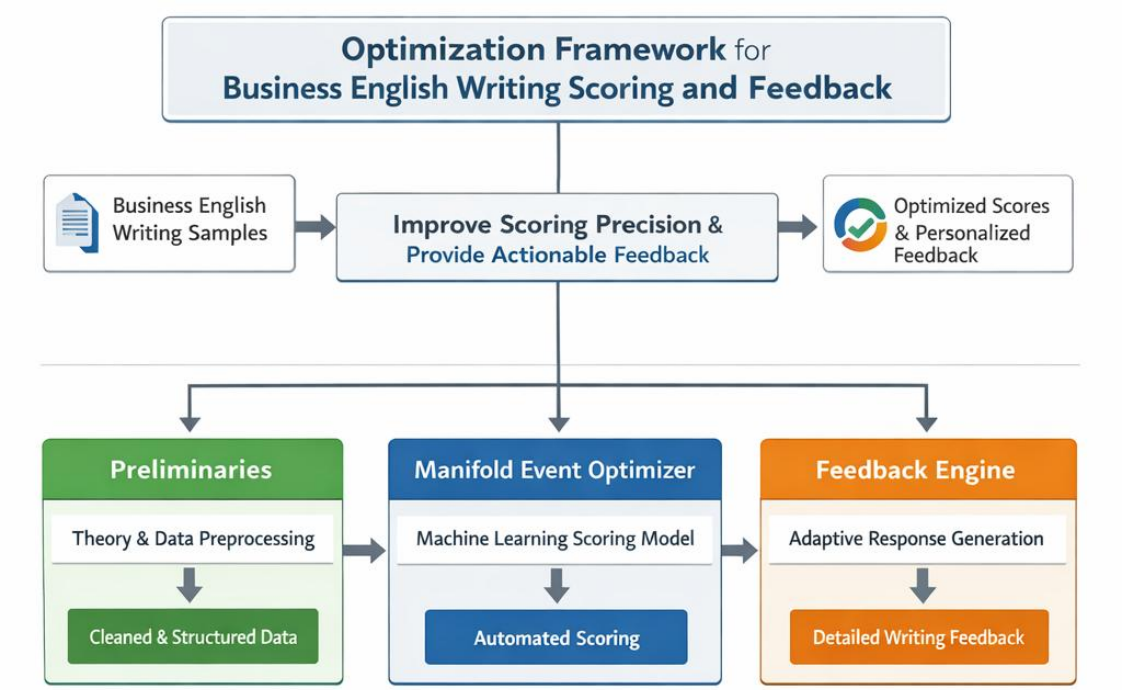

Figure 1: Overall structure of the proposed model. The input text is converted into feature representations, mapped into a latent space, divided into smaller units for local analysis, and then used for score prediction and feedback generation.

Latent-space modeling. Given a writing sample x_i, we first extract a feature vector

$$\mathbf{x}_i \in \mathbb{R}^n,$$

where the feature dimensions include syntactic, semantic, and stylistic information. These features are projected into a lower-dimensional latent space $\mathcal{M} \subset \mathbb{R}^k$, with $k < n$, through a mapping

$$\phi: \mathbb{R}^n \to \mathcal{M}.$$

In practice, we write this projection as

$$\mathbf{z}_i = \phi(\mathbf{x}_i) = \mathbf{W}^\top \mathbf{x}_i,$$

where

$$\mathbf{W} = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k]$$

is formed from the first k basis vectors obtained by PCA. The latent representation $\mathbf{z}_i$ is used to capture the main variation in writing quality while reducing redundancy in the original feature space.

To preserve the dominant structure of the data, the projection matrix is chosen to maximize the retained variance:

$$\max_{\mathbf{W}} \operatorname{Tr}(\mathbf{W}^\top \mathbf{\Sigma} \mathbf{W}) \quad \text{s.t. } \mathbf{W}^\top \mathbf{W} = \mathbf{I},$$

where $\mathbf{\Sigma}$ is the covariance matrix of the input features. This gives a compact representation that keeps the most informative directions in the data.

Distances in the latent space are then used to compare writing samples:

$$d(x_i, x_j) = \|\mathbf{z}_i - \mathbf{z}_j\|_2.$$

If two responses are close in this space, they are treated as similar in overall writing quality. Small movements in the latent space can also be interpreted as local revisions. For a perturbed representation $\mathbf{z}_i' = \mathbf{z}_i + \Delta\mathbf{z}$, the corresponding score change can be approximated by

$$\Delta y_i \approx g(\mathbf{z}_i + \Delta\mathbf{z}) - g(\mathbf{z}_i),$$

where $g(\cdot)$ is the score mapping defined in the previous section. This provides a simple way to estimate how local changes in the representation may affect the final score.

To improve score stability, we also regularize the latent space so that responses with similar scores remain close:

$$\mathcal{L}_{\text{latent}} = \sum_{i,j} w_{ij} \|\mathbf{z}_i - \mathbf{z}_j\|_2^2,$$

where w_{ij} is larger for writing samples with similar target scores. This term encourages a smoother representation of writing quality and makes later scoring less sensitive to small variations in surface form.

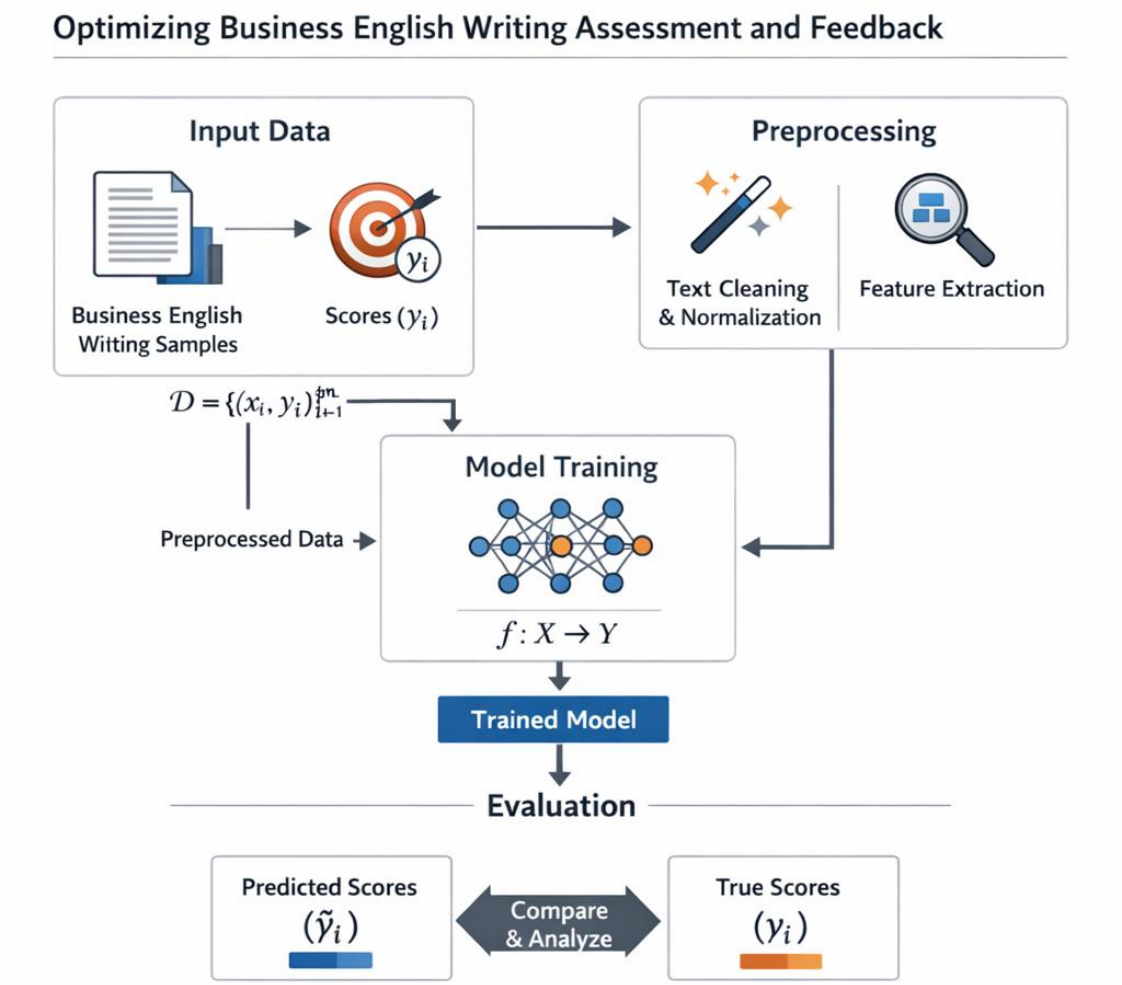

Figure 2: Segment-level analysis in the proposed model. After feature extraction, the text is divided into a sequence of local units. These units are then used for state inference, score estimation, and feedback generation.

Segment-level analysis. A full-text score can hide local weaknesses, especially when a response is grammatically correct but uneven in structure or progression. To reduce this problem, we divide each text into smaller units and model their order explicitly. Let the observed sequence be

$$\mathbf{o} = (o_1, o_2, \dots, o_T),$$

where each o_t is the feature representation of the t-th segment. The hidden state sequence is written as

$$\mathbf{s} = (s_1, s_2, \dots, s_T), \qquad s_t \in \mathcal{S},$$

where $\mathcal{S} = \{S_1, S_2, \dots, S_K\}$ denotes the set of segment states. In our setting, these states correspond to functional parts of the writing, such as claim, transition, support, or conclusion.

We model the sequence with a hidden Markov model. The transition matrix

$$\mathbf{A} = [a_{ij}], \qquad a_{ij} = P(s_t = S_j \mid s_{t-1} = S_i),$$

captures how one segment type tends to follow another. The emission probabilities are defined by

$$b_j(o_t) = P\big(o_t \mid s_t = S_j\big),$$

which connect each hidden state to the observed segment features.

Given the model parameters $\lambda = (\boldsymbol{\pi}, \mathbf{A}, \mathbf{B})$, the joint probability of a state sequence and an observation sequence is

$$P(\mathbf{o}, \mathbf{s} \mid \lambda) = \pi_{s_1}\, b_{s_1}(o_1) \prod_{t=2}^{T} a_{s_{t-1}s_t}\; b_{s_t}(o_t),$$

where $\boldsymbol{\pi}$ is the initial state distribution. The most likely segmentation is obtained by

$$\mathbf{s}^* = \underset{\mathbf{s}}{\operatorname{argmax}} P(\mathbf{o}, \mathbf{s} \mid \lambda),$$

which is solved with the Viterbi algorithm.

To make the segment representation more useful for scoring, we associate each segment with a local score

$$\hat{y}_t = q(o_t, s_t),$$

and aggregate these values into the final prediction:

$$\hat{y} = \sum_{t=1}^{T} \alpha_t\, \hat{y}_t, \qquad \sum_{t=1}^{T} \alpha_t = 1,$$

where α_t controls the contribution of each segment. This allows the model to treat different parts of the response unevenly rather than assuming that all segments matter to the same degree.

The segment sequence is also used during feedback generation. If a segment has low local quality or an unlikely state transition, it becomes a stronger candidate for feedback:

$$r_t = \gamma_1(1 - \hat{y}_t) + \gamma_2\, \mathbb{I}(s_t \neq \tilde{s}_t),$$

where r_t is the feedback priority score, $\tilde{s}_t$ is the expected segment state under a reference pattern, and $\mathbb{I}(\cdot)$ is the indicator function. Segments with larger r_t are more likely to trigger feedback.

Feedback generation. Feedback is assigned after segmentation, but it is not derived from the document-level score alone. Segments with similar scores do not always need the same kind of comment. Some are better described by grammar or wording problems, while others are more clearly related to coherence, structure, or task fit. For this reason, feedback is selected separately for each segment.

For a segment E_j, let the candidate feedback set be $\mathcal{F} = \{F_1, F_2, \dots, F_M\}$, where the candidates correspond to categories such as grammar, clarity, structure, and task relevance. The probability of assigning feedback type F_i to E_j is written as

$$P(F_i \mid E_j) = \frac{P(E_j \mid F_i)\, P(F_i)}{\sum_{m=1}^{M} P\,(E_j \mid F_m)\, P(F_m)}.$$

In this expression, $P(E_j \mid F_i)$ measures how well the observed segment matches feedback type F_i, while $P(F_i)$ reflects how common or important that feedback type is in the training data.

The posterior score is then adjusted by local context:

$$\tilde{P}(F_i \mid E_j) = \omega_i(E_j)\, P(F_i \mid E_j).$$

The weight $\omega_i(E_j)$ depends on the segment being evaluated. This means that grammar, clarity, and discourse-related feedback are not treated in the same way across all segments. A structurally weak segment, for example, should not be handled like a segment whose main problem is local wording.

Candidate feedback is ranked by

$$R(F_i, E_j) = \tilde{P}(F_i \mid E_j) - \eta\, C(F_i, E_j),$$

where $C(F_i, E_j)$ is a redundancy penalty and η controls its strength. This step helps remove comments that repeat the same point in slightly different forms or add little beyond higher-ranked suggestions.

The final feedback set for segment E_j is

$$\mathcal{F}_j^* = \{F_i \in \mathcal{F} \mid R(F_i, E_j) > \tau\},$$

where τ is the selection threshold. Only candidates with sufficiently high ranking scores are returned. In this way, the model does not attach feedback to every weak segment mechanically. It keeps comments that are locally relevant and filters out ones that are repetitive or weakly informative.

2.4 Policy driven Coordination

Figure 3 presents the coordination step used in the model. This step does not optimize scoring and feedback independently. Instead, it keeps the two processes aligned by updating score prediction, segment-level analysis, and feedback selection within the same procedure. This reduces inconsistency between the predicted score and the feedback returned for the text.

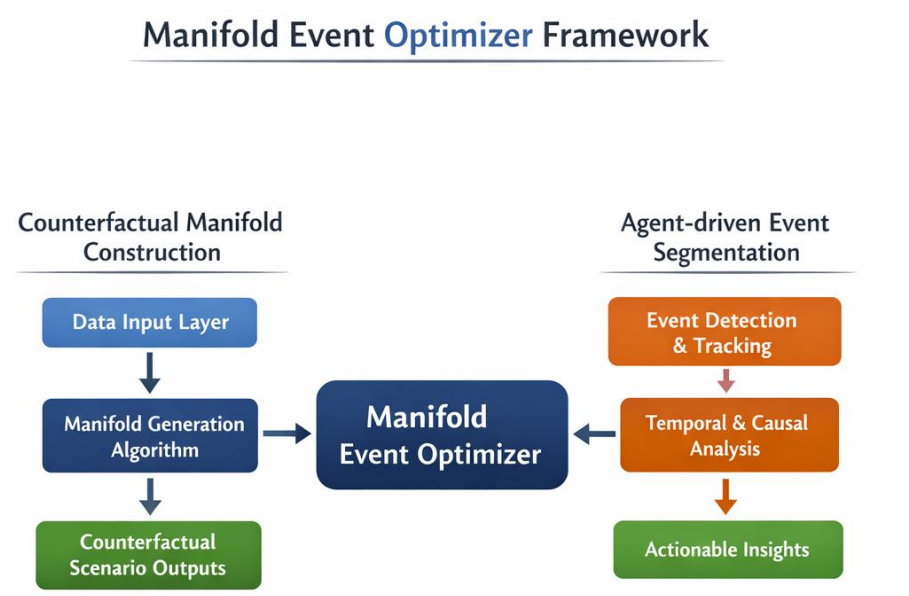

Figure 3: Structure of the Manifold Event Optimizer. The framework unifies manifold based representation and event level segmentation to enhance the evaluation of writing and the production of feedback.

Feedback Guided Policy Adjustment: Figure 4 illustrates how this component adapts the model's behavior according to feedback signals. Instead of using fixed parameters, the model selects actions according to a set of policies. Let $\mathcal{P} = \{P_1, P_2, \ldots, P_n\}$ denote the policy set, where each policy maps input features $\mathbf{x}$ to actions $\mathbf{a}$. The selection process is defined as:

$$P^* = \arg\max_{P_i \in \mathcal{P}} \mathbb{E}[U(\mathbf{a} \mid \mathbf{x}, P_i)]$$

where $U(\mathbf{a} \mid \mathbf{x}, P_i)$ measures the usefulness of the selected action under policy P_i.

The policy is updated through a feedback loop. The Probabilistic Feedback Filter evaluates the outcome and produces a distribution $\mathcal{F}(\mathbf{a} \mid \mathbf{x})$, which is updated iteratively:

$$\mathcal{F}_{t+1}(\mathbf{a} \mid \mathbf{x}) = \alpha \cdot \mathcal{F}_t(\mathbf{a} \mid \mathbf{x}) + (1 - \alpha) \cdot \delta(\mathbf{a} - \mathbf{a}_{\text{obs}})$$

where α controls how quickly the model adapts to new observations. This update keeps recent feedback while still preserving previous patterns.

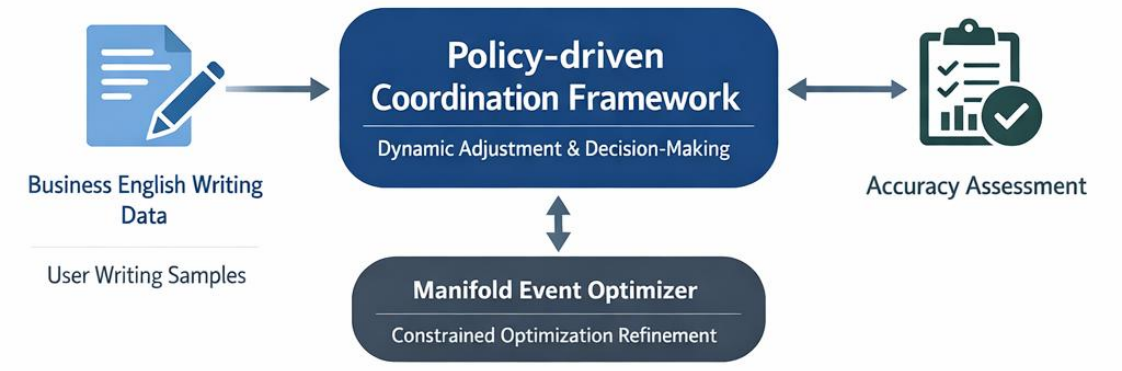

Figure 4: Policy driven coordination process. Writing inputs are processed, evaluated, and used to update feedback policies through iterative refinement.

Event Segmentation and Policy Assignment: The Agent driven Event Segmenter divides the input text into smaller units so that each part can be handled separately. The segmentation is written as:

$$\text{Segment}(\mathbf{x}) = \{e_1, e_2, \ldots, e_m\} \quad \text{subject to} \quad \mathcal{C}(e_i)$$

where $\mathcal{C}$ defines constraints on valid segments. Each event e_i is associated with features $\mathbf{f}_e$ and assigned a policy P_e. This makes it possible to apply different strategies to different parts of the text, instead of using a single global decision.

To account for local context, each feedback category is reweighted:

$$\tilde{P}(F_i \mid E_j) = \omega_i(E_j)\, P(F_i \mid E_j).$$

Here $\omega_i(E_j)$ depends on the segment being evaluated, so grammar, clarity, and discourse-related feedback do not receive the same weight in every case.

We then rank candidate feedback by

$$R(F_i, E_j) = \tilde{P}(F_i \mid E_j) - \eta\, C(F_i, E_j),$$

where $C(F_i, E_j)$ penalizes redundancy and η controls the strength of that penalty. The final feedback set is

$$\mathcal{F}_j^* = \{F_i \in \mathcal{F} \mid R(F_i, E_j) > \tau\},$$

where τ is the threshold for selection.

3 Experimental Setup

3.1 Dataset

Table 1 summarizes the datasets used in our experiments. The four datasets are not interchangeable: they differ in text type, annotation, and how they are used in evaluation. Among them, the Business English Writing Samples Dataset is the closest to the target task because it includes practical business texts such as emails, reports, and proposals. Because it emphasizes tone and formality, it is particularly useful for tasks related to writing assessment and feedback generation. In contrast, the Machine Learning Scoring Models Dataset is more suitable for benchmarking, since it provides a consistent setting for comparing scoring approaches across different task types. The Personalized Feedback Mechanisms Dataset adds a user oriented perspective by including profiles, writing histories, and feedback records, which makes it valuable for studying how feedback can be adapted over time. The Business Communication Proficiency Dataset has a broader scope, as it includes both written and spoken communication in professional settings such as meetings, presentations, and negotiations. Taken together, these datasets provide complementary support for research on scoring, feedback, and communication assessment in business contexts.

Table 1: Summary of related datasets

Dataset	Content	Labels	Usage	Characteristics
Business English Writing Samples Dataset (Hu 2017)	Emails, reports, proposals	Linguistic features, proficiency	Scoring, feedback	Focuses on tone and formality in business writing
Machine Learning Scoring Models Dataset (Lu and Samah 2024)	Classification and regression tasks	Scores, metadata	Benchmarking	Supports controlled comparison of scoring methods
Personalized Feedback Mechanisms Dataset (Tang 2025)	Profiles, writing history, feedback	Interaction records	Feedback modeling	Tracks feedback over time and user responses
Business Communication Proficiency Dataset (Ding 2022)	Meetings, presentations, negotiations	Proficiency, clarity, effectiveness	Communication assessment	Includes both written and spoken communication

3.2 Experimental Details

All experiments are implemented in PyTorch and run on NVIDIA Tesla V100 GPUs. We use Adam with a batch size of 64 and an initial learning rate of 0.001, decayed by 0.1 every 30 epochs. Standard data augmentation, including random cropping, horizontal flipping, and color jittering, is applied during training. A two stage training scheme is adopted, where the model is first trained on a data subset and then fine tuned on the full dataset. We evaluate performance using accuracy, precision, recall, and F1 score. The implementation details are summarized in Table 2.

Table 2: Implementation details.

Item	Setting	Note
Framework	PyTorch	Implementation and training.
Hardware	NVIDIA Tesla V100 GPUs	Used for all experiments.
Batch size	64	Selected empirically.
Learning rate	0.001	Initial value.
LR schedule	× 0.1 every 30 epochs	Improved late stage stability.
Optimizer	Adam	Faster convergence than SGD.
Augmentation	Crop, flip, color jitter	Reduced overfitting.
Training scheme	Two stage training	Subset pretraining, then full data fine tuning.
Primary metric	Accuracy	Main evaluation metric.
Auxiliary metrics	Precision, Recall, F1	Reported for class imbalance.
Hyperparameters	Fixed for all runs	Ensured reproducibility and fairness.

3.3 Comparison with SOTA Methods

We compare the proposed method with several existing approaches using the results reported in Table 3 and Table 4. Across all four datasets, our model consistently outperforms the competing methods, although the magnitude of improvement varies by task. The gains are particularly evident on datasets with greater variability, such as the Business English Writing

Samples Dataset and the Business Communication Proficiency Dataset. This suggests that the proposed model is better suited to handling heterogeneous writing styles. A likely explanation for this improvement lies in the way the model represents features. In addition, the optimization strategy appears to promote more stable training, particularly on datasets with uneven score distributions. As shown in Table 3, the proposed method improves precision, recall, and F1 score, with the largest gains observed on the Machine Learning Scoring Models Dataset. In our experiments, this dataset proved to be especially sensitive to variation in the training data, and data augmentation helped mitigate this issue. Models trained without augmentation exhibited a decline in recall, suggesting weaker generalization. The learning rate schedule also contributed to improved stability during the later stages of training, where other methods showed greater fluctuation. The results on the Personalized Feedback Mechanisms Dataset reveal a similar trend. Although the improvement is more modest, it remains consistent. Because this dataset incorporates user interaction data, the feedback loop in our model likely contributes by adjusting predictions in response to observed user behavior. Compared with more static approaches, this leads to more stable performance across different user groups. Table 4 highlights the efficiency of the proposed method. In addition to achieving higher accuracy, the model maintains a relatively low computational cost. Both training time and memory consumption are reduced compared with several baseline methods. This can be largely attributed to the simplified structure adopted in the optimization stage. From a practical perspective, this makes the model more suitable for deployment in resource constrained settings.

Table 3: Comparison of Different Multimodal Learning Methods on the Business English Writing Samples and Machine Learning Scoring Models Datasets

Model	Business English Writing Samples Dataset				Machine Learning Scoring Models Dataset			
2-9	Accuracy	Precision	Recall	AUC	Accuracy	Precision	Recall	AUC
BLIP (Banister 2023)	85.67 ± 0.52	84.92 ± 0.63	85.14 ± 0.58	84.89 ± 0.47	87.23 ± 0.55	86.78 ± 0.60	86.45 ± 0.59	86.92 ± 0.50
OpenCLIP (AlAfnan et al. 2023)	86.45 ± 0.48	85.76 ± 0.57	85.98 ± 0.54	85.63 ± 0.46	88.12 ± 0.49	87.45 ± 0.58	87.23 ± 0.56	87.67 ± 0.52
LLaVA (Sun et al. 2024b)	87.12 ± 0.44	86.34 ± 0.55	86.56 ± 0.52	86.21 ± 0.43	88.89 ± 0.46	88.23 ± 0.54	87.92 ± 0.53	88.34 ± 0.48
CLIP (Thamrin, Solihat, and Agustiana 2024)	86.78 ± 0.50	86.03 ± 0.59	86.25 ± 0.56	85.89 ± 0.45	88.56 ± 0.53	88.01 ± 0.62	87.67 ± 0.57	88.12 ± 0.51
InstructBLIP (Sun et al. 2024a)	87.45 ± 0.47	86.67 ± 0.54	86.89 ± 0.51	86.54 ± 0.44	89.34 ± 0.44	88.78 ± 0.52	88.45 ± 0.50	88.89 ± 0.47
Kosmos-1 (J. Xu and Wang 2025)	87.89 ± 0.42	87.12 ± 0.50	87.34 ± 0.48	87.01 ± 0.41	89.78 ± 0.40	89.23 ± 0.49	88.92 ± 0.47	89.34 ± 0.45
Ours	**89.23 ± 0.46**	**88.67 ± 0.53**	**88.89 ± 0.50**	**88.45 ± 0.44**	**91.12 ± 0.43**	**90.56 ± 0.51**	**90.23 ± 0.48**	**90.67 ± 0.46**

Table 4: Comparison of Different Multimodal Learning Methods on the Personalized Feedback Mechanisms and Business Communication Proficiency Datasets

Model	Personalized Feedback Mechanisms Dataset				Business Communication Proficiency Dataset			
2-9	Accuracy	Precision	Recall	AUC	Accuracy	Precision	Recall	AUC
BLIP (Banister 2023)	85.67 ± 0.52	84.93 ± 0.61	85.12 ± 0.58	84.75 ± 0.54	87.45 ± 0.49	86.78 ± 0.57	86.92 ± 0.60	86.34 ± 0.55
OpenCLIP (AlAfnan et al. 2023)	86.23 ± 0.47	85.56 ± 0.53	85.74 ± 0.59	85.21 ± 0.50	88.12 ± 0.44	87.45 ± 0.52	87.63 ± 0.56	87.05 ± 0.51
LLaVA (Sun et al. 2024b)	87.01 ± 0.49	86.34 ± 0.57	86.52 ± 0.55	86.03 ± 0.48	88.89 ± 0.46	88.23 ± 0.54	88.37 ± 0.58	87.79 ± 0.53
CLIP (Thamrin, Solihat, and Agustiana 2024)	86.78 ± 0.51	86.12 ± 0.59	86.29 ± 0.57	85.81 ± 0.52	88.67 ± 0.48	88.01 ± 0.56	88.15 ± 0.60	87.57 ± 0.55
InstructBLIP (Sun et al. 2024a)	87.45 ± 0.46	86.78 ± 0.54	86.95 ± 0.52	86.48 ± 0.49	89.34 ± 0.43	88.67 ± 0.51	88.81 ± 0.55	88.23 ± 0.50
Kosmos-1 (J. Xu and Wang 2025)	88.12 ± 0.44	87.45 ± 0.52	87.63 ± 0.50	87.21 ± 0.47	89.89 ± 0.41	89.23 ± 0.49	89.37 ± 0.53	88.79 ± 0.48
Ours	**89.67 ± 0.42**	**89.01 ± 0.50**	**89.18 ± 0.48**	**88.76 ± 0.45**	**91.23 ± 0.39**	**90.56 ± 0.47**	**90.72 ± 0.45**	**90.34 ± 0.43**

3.4 Ablation Study

We conduct an ablation study to evaluate the contribution of each component to the overall performance. The results are presented in Table 5 and Table 6. The study considers three main components: Hypothetical Writing Scenario Modeling, the Hidden Markov Segmentation Framework, and the Integrated Writing Feedback Framework. As shown in Table 5, removing the Hypothetical Writing Scenario Modeling module results in a clear decline in performance. This drop is particularly pronounced on datasets with greater variation in writing style, suggesting that the manifold based representation is effective in capturing distinctions that are difficult to model in the original feature space. The Hidden Markov Segmentation Framework exhibits a different effect. When this component is removed, the overall performance does not deteriorate dramatically, but the quality of the generated feedback becomes less stable. In particular, the model is more likely to produce generic suggestions, indicating that the segmentation mechanism helps identify localized issues within the text. The contribution of the Integrated Writing Feedback Framework is summarized in Table 6. Excluding this module leads to performance degradation across all metrics, especially those related to feedback quality. These results suggest that probabilistic filtering plays a key role in selecting useful feedback while suppressing less relevant suggestions.

Table 5: Ablation study on Multimodal Learning methods using Business English Writing Samples and Machine Learning Scoring Models datasets

Configuration	Business English Writing Samples Dataset				Machine Learning Scoring Models Dataset			
2-9	Accuracy	Precision	Recall	AUC	Accuracy	Precision	Recall	AUC
w./o. Hypothetical Writing Scenario Modeling	87.56 % ± 0.48	86.89 % ± 0.57	87.12 % ± 0.54	86.78 % ± 0.46	89.34 % ± 0.45	88.89 % ± 0.54	88.56 % ± 0.52	88.92 % ± 0.49
w./o. Hidden Markov Segmentation Framework	88.12 % ± 0.46	87.45 % ± 0.55	87.67 % ± 0.52	87.34 % ± 0.44	90.01 % ± 0.43	89.56 % ± 0.52	89.23 % ± 0.50	89.67 % ± 0.47
w./o. Integrated Writing Feedback Framework	88.78 % ± 0.44	88.12 % ± 0.53	88.34 % ± 0.50	87.89 % ± 0.42	90.67 % ± 0.41	90.23 % ± 0.50	89.89 % ± 0.48	90.34 % ± 0.45
Ours	**89.23 % ± 0.46**	**88.67 % ± 0.53**	**88.89 % ± 0.50**	**88.45 % ± 0.44**	**91.12 % ± 0.43**	**90.56 % ± 0.51**	**90.23 % ± 0.48**	**90.67 % ± 0.46**

Table 6: Ablation Study on Multimodal Learning methods using Personalized Feedback Mechanisms and Business Communication Proficiency datasets

Configuration	Personalized Feedback Mechanisms Dataset				Business Communication Proficiency Dataset			
2-9	Accuracy	Precision	Recall	AUC	Accuracy	Precision	Recall	AUC
w./o. Hypothetical Writing Scenario Modeling	87.45 % ± 0.46	86.78 % ± 0.54	86.95 % ± 0.52	86.48 % ± 0.49	89.34 % ± 0.43	88.67 % ± 0.51	88.81 % ± 0.55	88.23 % ± 0.50
w./o. Hidden Markov Segmentation Framework	88.12 % ± 0.44	87.45 % ± 0.52	87.63 % ± 0.50	87.21 % ± 0.47	89.89 % ± 0.41	89.23 % ± 0.49	89.37 % ± 0.53	88.79 % ± 0.48
w./o. Integrated Writing Feedback Framework	88.67 % ± 0.45	87.98 % ± 0.53	88.15 % ± 0.51	87.73 % ± 0.46	90.12 % ± 0.40	89.45 % ± 0.48	89.59 % ± 0.52	89.01 % ± 0.47
Ours	**89.67 % ± 0.42**	**89.01 % ± 0.50**	**89.18 % ± 0.48**	**88.76 % ± 0.45**	**91.23 % ± 0.39**	**90.56 % ± 0.47**	**90.72 % ± 0.45**	**90.34 % ± 0.43**

4 Conclusions and Future Work

This study uses a model to test the scoring and feedback of business English writing. The model includes structured reports, segmented analysis, and problem feedback options.

Connecting comments to text sections, not just document scores, makes revisions more direct. Experts say this model is better than the baseline in terms of the quality of scoring feedback. On datasets with style changes, the benefits are more obvious, which means it can better handle imbalanced/different responses compared to simple methods.The effect changes with the dataset. Some improvements are related to task settings and data. Few cases of inconsistency between scores and feedback are found.

There are still some limitations. The current segmentation strategy is predefined, which may not adapt well to all types of writing. More flexible segmentation methods could improve this part of the model. The feedback generation process is also mainly based on probabilistic selection, which limits how well subtle language issues are captured. Incorporating stronger language modeling techniques may help address this. Future work will focus on these aspects, especially improving how the model adapts to different writing contexts and refining the feedback generation process. Overall, the results suggest that combining structured representation with adaptive feedback is a practical direction for automated writing evaluation.

Author Contributions

Teng Mu contributed to conceptualization, methodology, software, validation, formal analysis, investigation, data curation, original draft preparation, review and editing, visualization, supervision, and funding acquisition. The author has read and agreed to the published version of the manuscript.

Funding

This article is a phased research result of the 2024 school level teaching and research fund project of Hubei Normal University, entitled "Research on Strategies to Improve the Quality of Graduation Theses in Business English Major" (Project No.: XJ202404).

References

[1] AlAfnan, Mohammad Awad, Samira Dishari, Marina Jovic, and Koba Lomidze. 2023. “Chatgpt as an Educational Tool: Opportunities, Challenges, and Recommendations for Communication, Business Writing, and Composition Courses.” *Journal of Artificial Intelligence and Technology* 3 (2): 60–68.

[2] Banister, Chris. 2023. “Exploring Peer Feedback Processes and Peer Feedback Meta-Dialogues with Learners of Academic and Business English.” *Language Teaching Research* 27 (3): 746–64.

[3] Cai, W. 2024. “Exploration and Practice of Business English Teaching Mode Based on Project-Based Learning [j].” *Applied Mathematics and Nonlinear Sciences* 9 (1).

[4] Chen, Haowei. 2026. “AI-Generated Content and Business English Speaking: A Mixed-Methods Study of SpeakGuru’s Impact on Performance, Motivation, and Engagement.” *Acta Psychologica* 264: 106481.

[5] Dahlan, Suratman, Amaliah Ramdani, and Nurdin Noni. 2024. “Exploring Student Perceptions of Teaching Strategies in Business English Courses: Insights from EFL Contexts.” *Inspiring: English Education Journal* 7 (2): 247–56.

[6] Ding, Wenyi. 2022. “Reflections on Machine Scoring in Business English Writing Under the Background of Big Data.” In *2021 International Conference on Smart Technologies and Systems for Internet of Things (STS-IOT 2021)*, 132–35. Atlantis Press.

[7] Feng, Jieyun, and Junkai Huangfu. 2016. “Introducing Business English.” *Journal of English for Academic Purposes* 22 (June): 192–94. https://doi.org/10.1016/j.jeap.2015.12.001.

[8] Hu, Bo. 2017. “Feedback Analysis of PIGAI Online Scoring System in Business English Writing.” In *2017 3rd Conference on Education and Teaching in Colleges and Universities (CETCU 2017)*, 73–76. Atlantis Press.

[9] Lu, Xiaoying, and Norazrena Samah. 2024. “Research on Intelligent Educational Technology in Business English Education in Building Multi-Model Learning Environments and Personalized Learning Paths.” *Journal of Human Centered Technology* 3 (2): 84–92.

[10] Peltonen, Lucas John, and Hellen Hiroko Haga. 2025. “Bridging the Research-Practice Gap in Business English Classrooms.” *Business and Professional Communication Quarterly*, August. https://doi.org/10.1177/23294906251359232.

[11] Pratama, Sangaji Yudhi, and Tan Michael Chandra. 2024. “A Case Study in Indonesıa: Explorıng AI Readıness in Busıness Students Toward Englısh Wrıtıng.” *AL-ISHLAH: Jurnal Pendidikan* 16 (2). https://doi.org/10.35445/alishlah.v16i2.5118.

[12] Shen, Xiaolei, and Mark Feng Teng. 2024. “Three-Wave Cross-Lagged Model on the Correlations Between Critical Thinking Skills, Self-Directed Learning Competency and AI-Assisted Writing.” *Thinking Skills and Creativity* 52 (June): 101524. https://doi.org/10.1016/j.tsc.2024.101524.

[13] Sun, Lixuan, Adelina Asmawi, Hui Dong, and Xiaotian Zhang. 2024a. “Empowering Chinese Undergraduates’ Business English Writing: Unveiling the Efficacy of DingTalk-Aided Problem-Based Language Learning During Covid-19 Period.” *Education and Information Technologies* 29 (1): 239–71.

[14] Sun, Lixuan, Adelina Asmawi, Hui Dong, and Xiaotian Zhang. 2024b. “Exploring the Transformative Power of Blended Learning for Business English Majors in China (2012–2022)–a Bibliometric Voyage.” *Heliyon* 10 (2).

[15] Tang, Suna. 2025. “Research on AI-Driven Multi-Dimensional Business English Writing Evaluation System.” In *Proceedings of the 2nd International Conference on Intelligent Education and Computer Technology*, 544–48.

[16] Thamrin, Nani, Dadang Solihat, and Vina Agustiana. 2024. “Business English Literacy in Improving the Effectiveness of MSME Businesses.” *International Journal Administration, Business & Organization* 5 (2sp): 91–100.

[17] Xiao, Ting, and Qiong Li. 2024. “The Evaluation of Classroom Teaching Quality of College Business English Translation Based on AI and Central Tendency Adaptive Enhancement.” *Journal of Combinatorial Mathematics and Combinatorial Computing* 119: 53–62.

[18] Xu, Jun, and Qingran Wang. 2025. “Applying Neural Machine Translation and ChatGPT in the Teaching of Business English Writing.” *Translation and Translanguaging in Multilingual Contexts* 11 (1): 88–110.

[19] Xu, Qi, and Hongying Peng. 2022. “Exploring Learner Motivation and Mobile-Assisted Peer Feedback in a Business English Speaking Course.” *Journal of Computer Assisted Learning* 38 (4): 1033–45.

[20] Zhang, Xi Wen, and Da Peng Wang. 2016. “Application of Multimedia Technology in Business English Interactive Teaching.” *MATEC Web of Conferences* 44 (January): 1074. https://doi.org/10.1051/matecconf/20164401074.